Experiment 4

Inverted index challenging exercise

Web scrapping documents and create Inverted Index after nessesary preprocessing

Name:	**:**	**Abhishek N N**	**Register No**	**:**	**20BCE1025**
Faculty	:	Dr. Alok Chauhan	Slot	:	L51-L52 AB1-605B
Course	:	Web Mining Lab	Code	:	CSE3024
Programme	:	B.Tech CSE Core	Semester	:	Win – 22 - 23

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

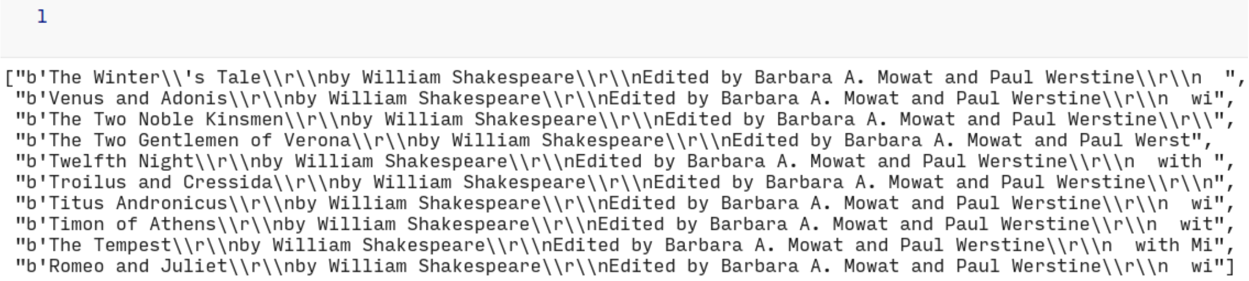
Problem- 1:

Collect any 10 documents (English text documents) from the web and create inverted index by doing necessary preprocessing steps using python.

```
# preprocessing
# import the necessary libraries
import string
import nltk
import re
def text_lowercase(text):
    return text.lower()
# Remove numbers
def remove_numbers(text):
    result = re.sub(r'\d+', '', text)
    return result
# remove punctuation
def remove_punctuation(text):
    translator = str.maketrans('', '', string.punctuation)
    return text.translate(translator)
# remove whitespace from text
def remove_whitespace(text):
    return  " ".join(text.split())
# remove stopwords
def remove_stopwords(text):
    stopword_list = nltk.corpus.stopwords.words('english')
    tokens = nltk.word_tokenize(text)
    tokens = [token.strip() for token in tokens]
    return ' '.join([token for token in tokens if token not
in stopword_list])
# stemming
def stem_text(text):
    ps = nltk.PorterStemmer()
    tokens = nltk.word_tokenize(text)
    tokens = [token.strip() for token in tokens]
    return ' '.join([ps.stem(token) for token in tokens])
# lemmatization
def lemmatize_text(text):
    wnl = nltk.WordNetLemmatizer()
    tokens = nltk.word_tokenize(text)
    tokens = [token.strip() for token in tokens]
    return ' '.join([wnl.lemmatize(token) for token in
tokens])
# remove special characters
def remove_special_characters(text):
    pattern=r'[^a-zA-z0-9\s]'
    text=re.sub(pattern,'',text)
    return text
# remove extra newlines
def remove_extra_newlines(text):
    pattern=r'[\r|\n|\r]+'
    text=re.sub(pattern,' ',text)
    return text

# apply all the functions to the text
def preprocess(corpus):
    normalized_corpus = []
    # normalize each document in the corpus
    for doc in corpus:
        doc = text_lowercase(doc)
        doc = remove_numbers(doc)
        doc = remove_punctuation(doc)
        doc = remove_whitespace(doc)
        doc = remove_special_characters(doc)
        doc = remove_extra_newlines(doc)
        doc = lemmatize_text(doc)
        doc = stem_text(doc)
        doc = remove_stopwords(doc)
        normalized_corpus.append(doc)
    return normalized_corpus
```

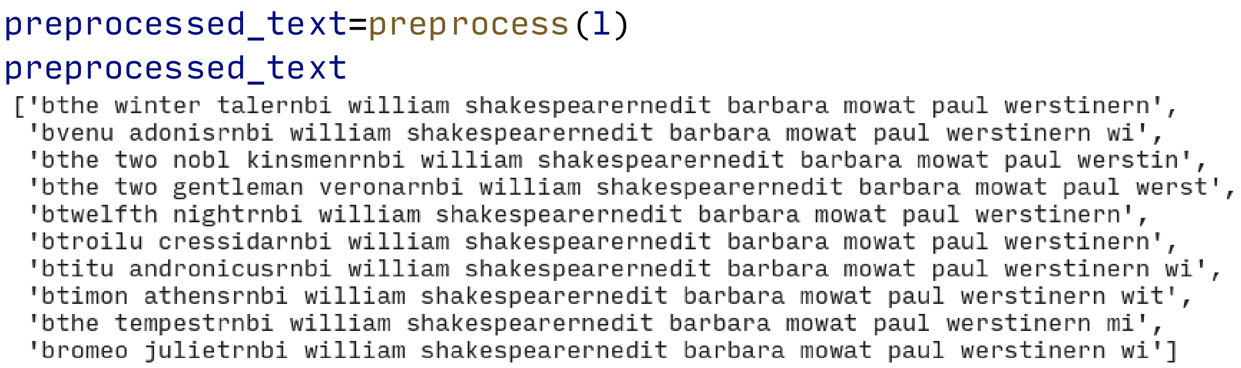
```
# inverted index
def generateInvertedIndexDict(dataFromDoc: list[str]) :
    d=dict()
    termsListFromDoc = [s.split() for s in dataFromDoc]

    for docId, termList in enumerate(termsListFromDoc):
        for term in termList:
            if term not in d:
                d[term]={docId}
            else:
                d[term].add(docId)
    return d

# file handling
from os import listdir
from os.path import isfile, join

def getDataFromDocs(dir):
    """
    gets strings from docs
    parameters:
    dir (str) : directroy which contains all files
    return:
    list of str read from docs in the directory given by user
    """
    return [open(join(dir, f)).read() for f in
sorted(listdir(dir)) if isfile(join(dir, f))]

def getDocIDToDocNameMap(dir):
    """
    gets the map of docID to docName
    parameters:
    dir (str) : directroy which contains all files
    return:
    dict of docID to docName of docs in the directory given
by user
    """
    return {i:x for i, x in enumerate([f for f in
sorted(listdir(dir)) if isfile(join(dir, f))])}
```

```python
from urllib.request import urlopen
urls=[
    "https://shakespeare.folger.edu/downloads/txt/the-winters-tale_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/venus-and-adonis_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/the-two-noble-kinsmen_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/the-two-gentlemen-of-verona_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/twelfth-night_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/troilus-and-cressida_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/titus-andronicus_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/timon-of-athens_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/the-tempest_TXT_FolgerShakespeare.txt",
    "https://shakespeare.folger.edu/downloads/txt/romeo-and-juliet_TXT_FolgerShakespeare.txt"
      ]
l=[]
for url in urls:
    textPage = urlopen(url)
    l.append(textPage.read())

# inverted index becomes to long so took only first 100 characters
for i in range(len(l)):
    l[i]=str(l[i])[:100]
```

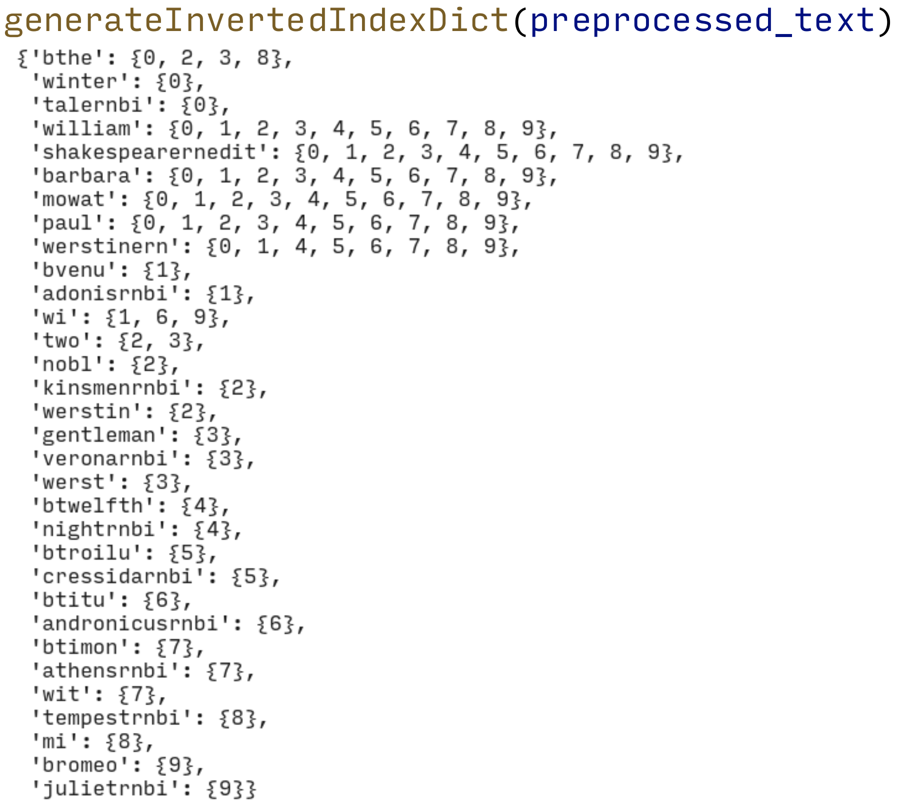


```
l
```

```
["b'The Winter\\'s Tale\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  ",
 "b'Venus and Adonis\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  wi",
 "b'The Two Noble Kinsmen\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\",
 "b'The Two Gentlemen of Verona\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werst",
 "b'Twelfth Night\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  with ",
 "b'Troilus and Cressida\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n",
 "b'Titus Andronicus\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  wi",
 "b'Timon of Athens\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  wit",
 "b'The Tempest\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  with Mi",
 "b'Romeo and Juliet\\r\\nby William Shakespeare\\r\\nEdited by Barbara A. Mowat and Paul Werstine\\r\\n  wi"]
```



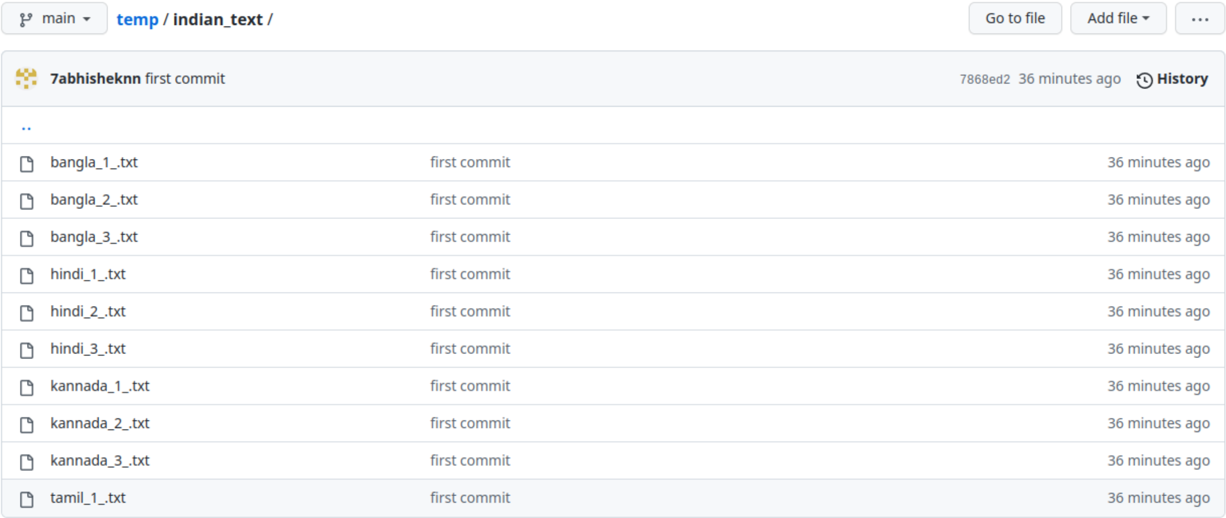
```
preprocessed_text=preprocess(l)
preprocessed_text
['bthe winter talernbi william shakespearernedit barbara mowat paul werstinern',
 'bvenu adonisrnbi william shakespearernedit barbara mowat paul werstinern wi',
 'bthe two nobl kinsmenrnbi william shakespearernedit barbara mowat paul werstin',
 'bthe two gentleman veronarnbi william shakespearernedit barbara mowat paul werst',
 'btwelfth nightrnbi william shakespearernedit barbara mowat paul werstinern',
 'btroilu cressidarnbi william shakespearernedit barbara mowat paul werstinern',
 'btitu andronicusrnbi william shakespearernedit barbara mowat paul werstinern wi',
 'btimon athensrnbi william shakespearernedit barbara mowat paul werstinern wit',
 'bthe tempestrnbi william shakespearernedit barbara mowat paul werstinern mi',
 'bromeo julietrnbi william shakespearernedit barbara mowat paul werstinern wi']
```



```
generateInvertedIndexDict(preprocessed_text)
{'bthe': {0, 2, 3, 8},
 'winter': {0},
 'talernbi': {0},
 'william': {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},
 'shakespearernedit': {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},
 'barbara': {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},
 'mowat': {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},
 'paul': {0, 1, 2, 3, 4, 5, 6, 7, 8, 9},
 'werstinern': {0, 1, 4, 5, 6, 7, 8, 9},
 'bvenu': {1},
 'adonisrnbi': {1},
 'wi': {1, 6, 9},
 'two': {2, 3},
 'nobl': {2},
 'kinsmenrnbi': {2},
 'werstin': {2},
 'gentleman': {3},
 'veronarnbi': {3},
 'werst': {3},
 'btwelfth': {4},
 'nightrnbi': {4},
 'btroilu': {5},
 'cressidarnbi': {5},
 'btitu': {6},
 'andronicusrnbi': {6},
 'btimon': {7},
 'athensrnbi': {7},
 'wit': {7},
 'tempestrnbi': {8},
 'mi': {8},
 'bromeo': {9},
 'julietrnbi': {9}}
```

Challenging Exercise-1:

Collect any 10 documents (Indian Language text Documents in Unicode) from the web and create inverted index by doing necessary preprocessing steps using python.



temp / indian_text /

7abhisheknn first commit 7868ed2 36 minutes ago

..

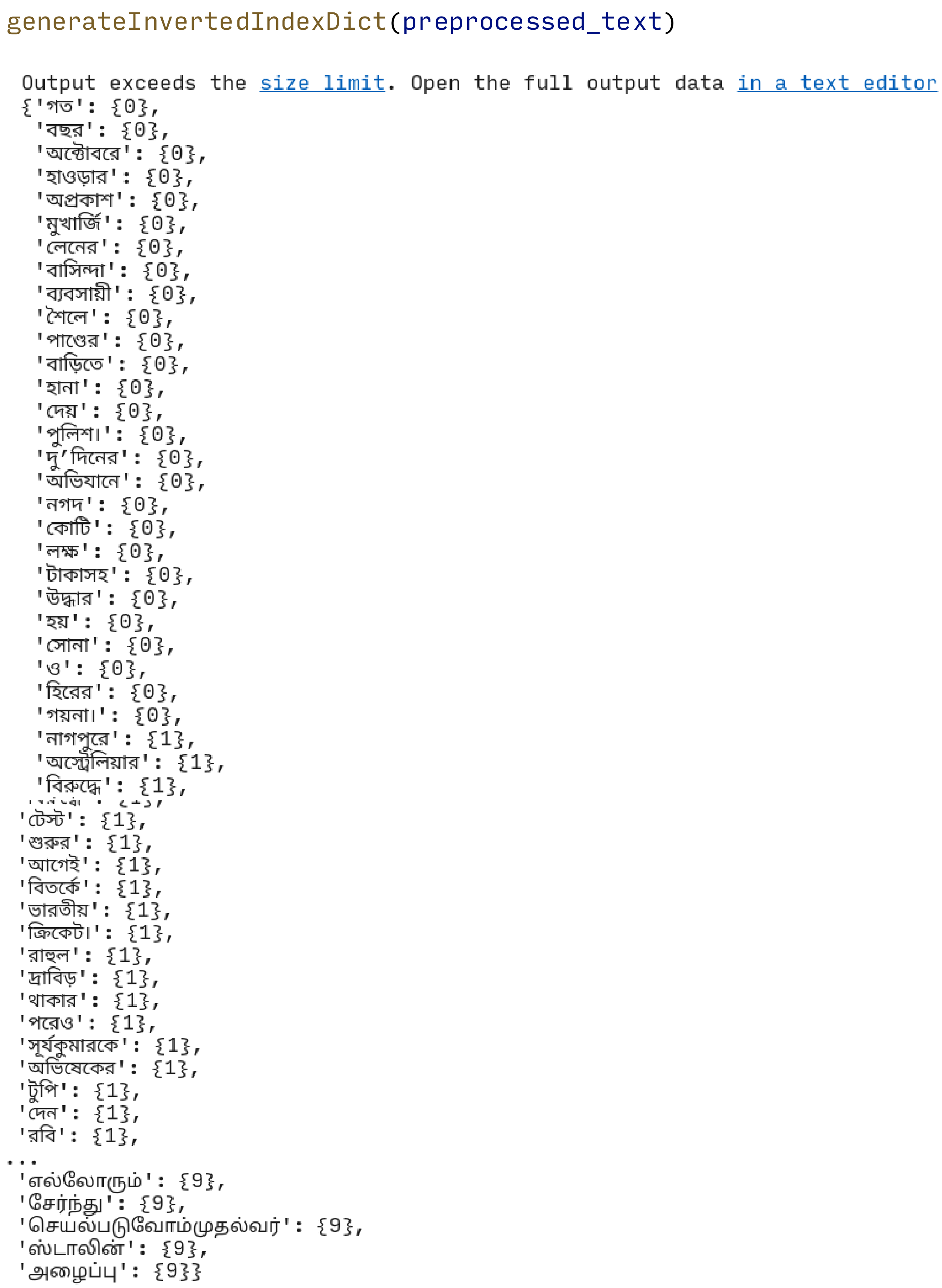
bangla_1_.txt	first commit	36 minutes ago
bangla_2_.txt	first commit	36 minutes ago
bangla_3_.txt	first commit	36 minutes ago
hindi_1_.txt	first commit	36 minutes ago
hindi_2_.txt	first commit	36 minutes ago
hindi_3_.txt	first commit	36 minutes ago
kannada_1_.txt	first commit	36 minutes ago
kannada_2_.txt	first commit	36 minutes ago
kannada_3_.txt	first commit	36 minutes ago
tamil_1_.txt	first commit	36 minutes ago

```python
# direct web link to indian text file was not there so i made github folder containing files
urls=[

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indian_text/bangla_1_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indian_text/bangla_2_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indian_text/bangla_3_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indian_text/hindi_1_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indian_text/hindi_2_.txt",
"https://raw.githubusercontent.com/7abhisheknn/temp/main/indi
an_text/hindi_3_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indi
an_text/kannada_1_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indi
an_text/kannada_2_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indi
an_text/kannada_3_.txt",

"https://raw.githubusercontent.com/7abhisheknn/temp/main/indi
an_text/tamil_1_.txt",
      ]
l=[]
for url in urls:
    textPage = urlopen(url)
    l.append(textPage.read().decode('utf-8'))
l
```

```
['গত বছর অক্টোবরে হাওড়ার অপ্রকাশ মুখার্জি লেনের বাসিন্দা ব্যবসায়ী শৈলে পাণ্ডের বাড়িতে হানা দেয় পুলিশ। দু’দিনের অভিযানে নগদ
৮ কোটি ১৫ লক্ষ টাকা-সহ উদ্ধার হয় সোনা ও হিরের গয়না। ',
 'নাগপুরে অস্ট্রেলিয়ার বিরুদ্ধে টেস্ট শুরুর আগেই বিতর্কে ভারতীয় ক্রিকেট। রাহুল দ্রাবিড় থাকার পরেও সূর্যকুমারকে অভিষেকের টুপি দেন
রবি শাস্ত্রী। এই ঘটনা নিয়েই প্রশ্ন উঠছে। ',
 'রত্নগিরির পেট্রো রসায়নের অন্দরের খবর এবং এই প্রকল্প ঘিরে যে দুষ্টচক্র গড়ে উঠেছে, তা নিয়ে সোমবার সংবাদপত্রে লিখেছিলেন
শশীকান্ত। তার পরই মঙ্গলবার গাড়িচাপা দিয়ে খুন করা হয়। ',
 "महाविनाश के बीच तुर्की में 'भूकंप टैक्स' पर आक्रोश, हजारों की मौत के बाद फूटा लोगों का गुस्सा ",
 'राहुल के 51 मिनट Vs पीएम मोदी के 88 मिनट, किसके भाषण में कौन से मुद्दे रहे हावी?',
 'बॉर्डर-गावस्कर ट्रॉफी कल से...पहला मुकाबला नागपुर में:इंडिया के टॉप ऑर्डर और स्पिनर्स का रोल अहम, वे 5 फैक्टर जो सीरीज का रिजल्ट
तय करेंगे',
 'ರಾಮಕೃಷ್ಣ ಹೆಗಡೆಯವರಿಗೆ ಕಲ್ಲು ಹೊಡೆದವರು ಯಾರು? ರಕ್ತದಲ್ಲೇ ಬ್ರಾಹ್ಮಣ ವಿರೋಧಿತನ ಇದೆ; 3 ಜಿಲ್ಲೆ ಇಟ್ಟುಕೊಂಡು ಸಿಎಂ ಆಗುವ ಕನಸೇತಕೆ?',
 'ಜನವರಿಯಲ್ಲಿ  ಟಾಟಾ ಮೋಟಾರ್ಸ್ 47,987 ಕಾರುಗಳನ್ನು ಮಾರಾಟ ಮಾಡಿದೆ. ವಾರ್ಷಿಕ ಆಧಾರದ ಮೇಲೆ ನೋಡುವುದಾದರೆ ಕಾರಿನ ಮಾರಾಟ
17.68 ಪ್ರತಿಶತದಷ್ಟು ಹೆಚ್ಚಾಗಿದೆ. ಕಳೆದ ವರ್ಷ ಈ ಜನವರಿಯಲ್ಲಿ ಕೇವಲ 40777 ಯೂನಿಟ್\u200cಗಳಷ್ಟು ಟಾಟಾ ಕಾರು ಮಾರಾಟವಾಗಿತ್ತು.
ಆದರೆ, ಜನವರಿ 2023 ರಲ್ಲಿ, ಕಂಪನಿಯು ಇದಕ್ಕಿಂತ 7210 ಹೆಚ್ಚು ಯುನಿಟ್\u200cಗಳನ್ನು ಕಂಪನಿ ಮಾರಾಟ ಮಾಡಿದೆ. ಇದರೊಂದಿಗೆ, ಟಾಟಾ
ಮೋಟಾರ್ಸ್ ದೇಶದ ಮೂರನೇ ಅತಿದೊಡ್ಡ ಕಾರು ಮಾರಾಟ ಕಂಪನಿಯಾಗಿ ಹೊರ ಹೊಮ್ಮಿದೆ. ಕಂಪನಿಯ ಹೆಚ್ಚು ಮಾರಾಟವಾಗುವ ಕಾರುಗಳು ಯಾವುವು
ನೋಡೋಣ. ',
 'ನವದೆಹಲಿ: ಭೂಕಂಪದಿಂದ ಜರ್ಝರಿತವಾಗಿರುವ ಟರ್ಕಿ ಮತ್ತು ಸಿರಿಯಾ (Turkey and Syria Earthquake) ದೇಶಗಳಿಗೆ ಜಗತ್ತಿನ ಅನೇಕ
ದೇಶಗಳು ಸಹಾಯಹಸ್ತ ಚಾಚಿವೆ. ಭಾರತ ಕೂಡ ವಿವಿಧ ರಕ್ಷಣಾ ಸಾಮಗ್ರಿಗಳನ್ನು (India Rescue Team To Turkey) ಟರ್ಕಿಗೆ
ಕಳುಹಿಸಿಕೊಟ್ಟಿದೆ. ಔಷಧಿ, ರಕ್ಷಣಾ ಸಿಬ್ಬಂದಿ, ಶ್ವಾನ ದಳ ಇತ್ಯಾದಿ ನೆರವನ್ನು ಭಾರತ ಒದಗಿಸುತ್ತಿದೆ. ಈಗಾಗಲೇ ಟರ್ಕಿಗೆ ಭಾರತದಿಂದ ನಾಲ್ಕು ಮಿಲಿಟರಿ
ವಿಮಾನಗಳು ಹೋಗಿವೆ. ಆದರೆ ಟರ್ಕಿಗೆ ಹೋಗುತ್ತಿದ್ದ ಭಾರತೀಯ ವಿಮಾನಗಳನ್ನು ಪಾಕಿಸ್ತಾನ ತಡೆದಿವೆ ಎನ್ನುವಂತಹ ಸುದ್ದಿಗಳು ನಿನ್ನೆ ಮಂಗಳವಾರ
ಸೋಷಿಯಲ್ ಮೀಡಿಯಾಗಳಲ್ಲಿ (Social Media) ವೈರಲ್ ಆಗಿದ್ದವು. ಟರ್ಕಿಗೆ ಅದರ ಮಿತ್ರದೇಶವೇ ಅಡ್ಡಿಪಡಿಸುತ್ತಿದೆ ಎನ್ನುವಂತಹ ವಿಮರ್ಶೆಗಳು
ವ್ಯಕ್ತವಾಗಿದ್ದವು.',
 '8 கோடி மக்களும் வாழ்த்த வேண்டும்.. எல்லோரும் சேர்ந்து செயல்படுவோம்..முதல்வர் ஸ்டாலின்
அழைப்பு']
```

```python
# since the data is indian we cannot apply all preprocessing
steps

def preprocess_indian_text(corpus):
    normalized_corpus = []
    # normalize each document in the corpus
    for doc in corpus:
        doc = remove_numbers(doc)
        doc = remove_punctuation(doc)
        doc = remove_whitespace(doc)
        doc = remove_extra_newlines(doc)
        normalized_corpus.append(doc)
    return normalized_corpus

preprocessed_text=preprocess_indian_text(l)
preprocessed_text
```

```
['গত বছর অক্টোবরে হাওড়ার অপ্রকাশ মুখার্জি লেনের বাসিন্দা ব্যবসায়ী শৈলে পাণ্ডের বাড়িতে হানা দেয় পুলিশ। দু’দিনের অভিযানে নগদ কোটি
লক্ষ টাকাসহ উদ্ধার হয় সোনা ও হিরের গয়না।',
 'নাগপুরে অস্ট্রেলিয়ার বিরুদ্ধে টেস্ট শুরুর আগেই বিতর্কে ভারতীয় ক্রিকেট। রাহুল দ্রাবিড় থাকার পরেও সূর্যকুমারকে অভিষেকের টুপি দেন রবি
শাস্ত্রী। এই ঘটনা নিয়েই প্রশ্ন উঠছে।',
 'রত্নগিরির পেট্রো রসায়নের অন্দরের খবর এবং এই প্রকল্প ঘিরে যে দুষ্টচক্র গড়ে উঠেছে তা নিয়ে সোমবার সংবাদপত্রে লিখেছিলেন শশীকান্ত।
তার পরই মঙ্গলবার গাড়িচাপা দিয়ে খুন করা হয়।',
 'महाविनाश के बीच तुर्की में भूकंप टैक्स पर आक्रोश हजारों की मौत के बाद फूटा लोगों का गुस्सा',
 'राहुल के मिनट Vs पीएम मोदी के मिनट किसके भाषण में कौन से मुद्दे रहे हावी',
 'बॉर्डरगावस्कर ट्रॉफी कल सेपहला मुकाबला नागपुर मेंइंडिया के टॉप ऑर्डर और स्पिनर्स का रोल अहम वे फैक्टर जो सीरीज का रिजल्ट तय करेंगे',
 'ರಾಮಕೃಷ್ಣ ಹೆಗಡೆಯವರಿಗೆ ಕಲ್ಲು ಹೊಡೆದವರು ಯಾರು ರಕ್ತದಲ್ಲೇ ಬ್ರಾಹ್ಮಣ ವಿರೋಧಿತನ ಇದೆ ಜಿಲ್ಲೆ ಇಟ್ಟುಕೊಂಡು ಸಿಎಂ ಆಗುವ ಕನಸೇತಕೆ',
 'ಜನವರಿಯಲ್ಲಿ ಟಾಟಾ ಮೋಟಾರ್ಸ್ ಕಾರುಗಳನ್ನು ಮಾರಾಟ ಮಾಡಿದೆ ವಾರ್ಷಿಕ ಆಧಾರದ ಮೇಲೆ ನೋಡುವುದಾದರೆ ಕಾರಿನ ಮಾರಾಟ ಪ್ರತಿಶತದಷ್ಟು ಹೆಚ್ಚಾಗಿದೆ
ಕಳೆದ ವರ್ಷ ಈ ಜನವರಿಯಲ್ಲಿ ಕೇವಲ ಯೂನಿಟ್\u200cಗಳಷ್ಟು ಟಾಟಾ ಕಾರು ಮಾರಾಟವಾಗಿತ್ತು ಆದರೆ ಜನವರಿ ರಲ್ಲಿ ಕಂಪನಿಯು ಇದಕ್ಕಿಂತ ಹೆಚ್ಚು
ಯುನಿಟ್\u200cಗಳನ್ನು ಕಂಪನಿ ಮಾರಾಟ ಮಾಡಿದೆ ಇದರೊಂದಿಗೆ ಟಾಟಾ ಮೋಟಾರ್ಸ್ ದೇಶದ ಮೂರನೇ ಅತಿದೊಡ್ಡ ಕಾರು ಮಾರಾಟ ಕಂಪನಿಯಾಗಿ ಹೊರ
ಹೊಮ್ಮಿದೆ ಕಂಪನಿಯ ಹೆಚ್ಚು ಮಾರಾಟವಾಗುವ ಕಾರುಗಳು ಯಾವುವು ನೋಡೋಣ',
 'ನವದೆಹಲಿ ಭೂಕಂಪದಿಂದ ಜರ್ಝರಿತವಾಗಿರುವ ಟರ್ಕಿ ಮತ್ತು ಸಿರಿಯಾ Turkey and Syria Earthquake ದೇಶಗಳಿಗೆ ಜಗತ್ತಿನ ಅನೇಕ ದೇಶಗಳು
ಸಹಾಯಹಸ್ತ ಚಾಚಿವೆ ಭಾರತ ಕೂಡ ವಿವಿಧ ರಕ್ಷಣಾ ಸಾಮಗ್ರಿಗಳನ್ನು India Rescue Team To Turkey ಟರ್ಕಿಗೆ ಕಳುಹಿಸಿಕೊಟ್ಟಿದೆ ಔಷಧಿ ರಕ್ಷಣಾ
ಸಿಬ್ಬಂದಿ ಶ್ವಾನ ದಳ ಇತ್ಯಾದಿ ನೆರವನ್ನು ಭಾರತ ಒದಗಿಸುತ್ತಿದೆ ಈಗಾಗಲೇ ಟರ್ಕಿಗೆ ಭಾರತದಿಂದ ನಾಲ್ಕು ಮಿಲಿಟರಿ ವಿಮಾನಗಳು ಹೋಗಿವೆ ಆದರೆ ಟರ್ಕಿಗೆ
ಹೋಗುತ್ತಿದ್ದ ಭಾರತೀಯ ವಿಮಾನಗಳನ್ನು ಪಾಕಿಸ್ತಾನ ತಡೆದಿವೆ ಎನ್ನುವಂತಹ ಸುದ್ದಿಗಳು ನಿನ್ನೆ ಮಂಗಳವಾರ ಸೋಷಿಯಲ್ ಮೀಡಿಯಾಗಳಲ್ಲಿ Social Media
ವೈರಲ್ ಆಗಿದ್ದವು ಟರ್ಕಿಗೆ ಅದರ ಮಿತ್ರದೇಶವೇ ಅಡ್ಡಿಪಡಿಸುತ್ತಿದೆ ಎನ್ನುವಂತಹ ವಿಮರ್ಶೆಗಳು ವ್ಯಕ್ತವಾಗಿದ್ದವು',
 'கோடி மக்களும் வாழ்த்த வேண்டும் எல்லோரும் சேர்ந்து செயல்படுவோம்முதல்வர் ஸ்டாலின் அழைப்பு']
```



```
generateInvertedIndexDict(preprocessed_text)
```

```
 Output exceeds the size limit. Open the full output data in a text editor
 {'গত': {0},
  'বছর': {0},
  'অক্টোবরে': {0},
  'হাওড়ার': {0},
  'অপ্রকাশ': {0},
  'মুখার্জি': {0},
  'লেনের': {0},
  'বাসিন্দা': {0},
  'ব্যবসায়ী': {0},
  'শৈলে': {0},
  'পাণ্ডের': {0},
  'বাড়িতে': {0},
  'হানা': {0},
  'দেয়': {0},
  'পুলিশ।': {0},
  'দু’দিনের': {0},
  'অভিযানে': {0},
  'নগদ': {0},
  'কোটি': {0},
  'লক্ষ': {0},
  'টাকাসহ': {0},
  'উদ্ধার': {0},
  'হয়': {0},
  'সোনা': {0},
  'ও': {0},
  'হিরের': {0},
  'গয়না।': {0},
  'নাগপুরে': {1},
  'অস্ট্রেলিয়ার': {1},
  'বিরুদ্ধে': {1},
 'টেস্ট': {1},
 'শুরুর': {1},
 'আগেই': {1},
 'বিতর্কে': {1},
 'ভারতীয়': {1},
 'ক্রিকেট।': {1},
 'রাহুল': {1},
 'দ্রাবিড়': {1},
 'থাকার': {1},
 'পরেও': {1},
 'সূর্যকুমারকে': {1},
 'অভিষেকের': {1},
 'টুপি': {1},
 'দেন': {1},
 'রবি': {1},
...
 'எல்லோரும்': {9},
 'சேர்ந்து': {9},
 'செயல்படுவோம்முதல்வர்': {9},
 'ஸ்டாலின்': {9},
 'அழைப்பு': {9}}
```

Challenging Exercise-2:

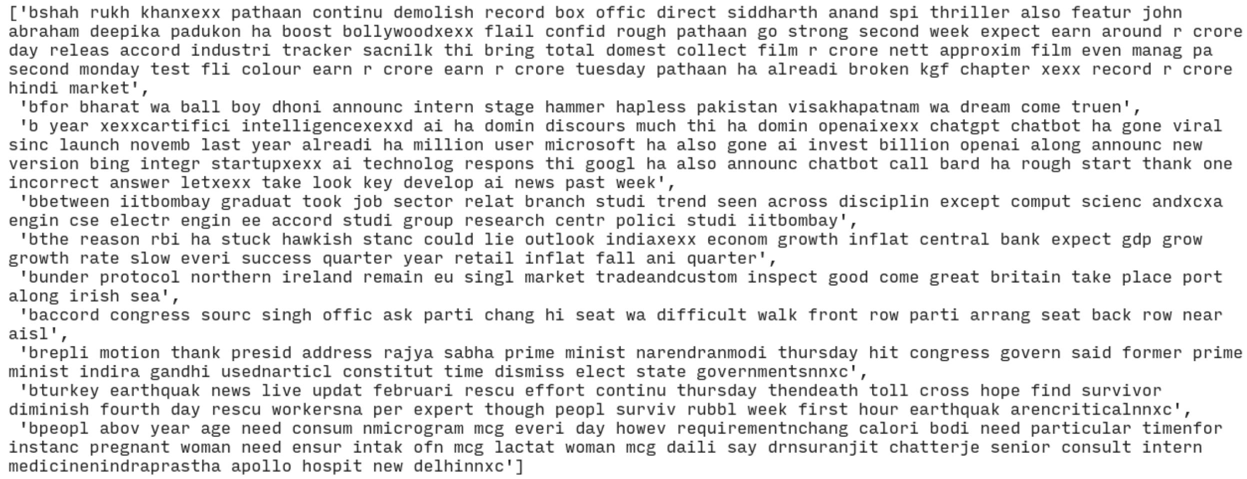
Collect any 10 documents (Documents in different formats such as PDF, DOC, ODF) from the web and create inverted index by doing necessary preprocessing steps using python.

```
import requests
import textract
urls=[

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/1.odt",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/2.odt",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/3.odt",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/4.odt",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/1.docx",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/2.docx",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/3.docx",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/1.pdf",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/2.pdf",

"https://github.com/7abhisheknn/temp/raw/main/different_format_documents/3.pdf",
      ]
```

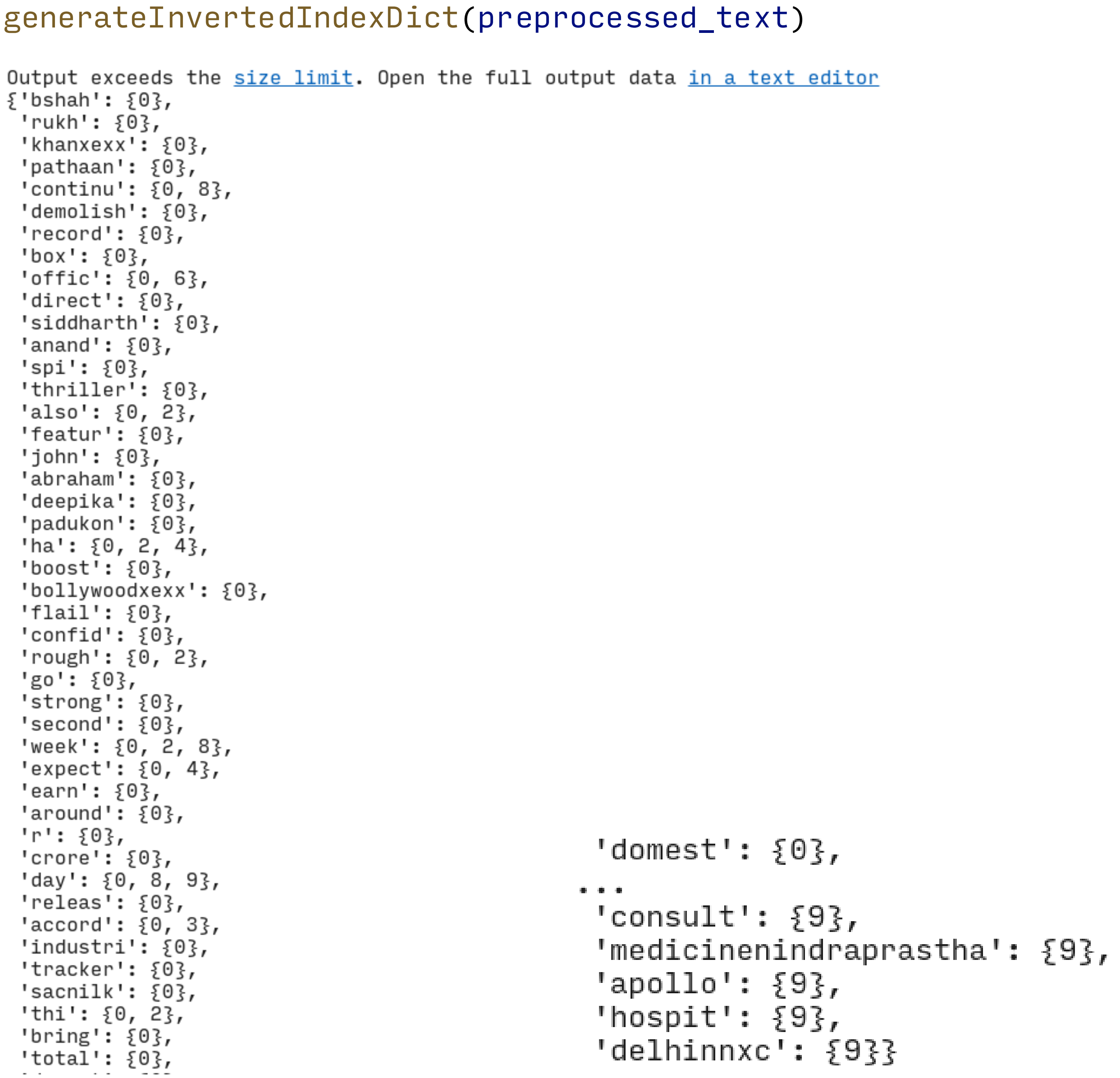
```
l=[]
for url in urls:
  response = requests.get(url)

  saveFile=""
  if (url[-3:]=="odt"):
    saveFile="temp.odt"
  elif (url[-3:]=="ocx"):
    saveFile="temp.docx"
  else:
    saveFile="temp.pdf"

  open(saveFile, "wb").write(response.content)
  text = textract.process(saveFile)
  l.append(text)
l
```

```
[b'Shah Rukh Khan\xe2\x80\x99s Pathaan continues to demolish records at the box office. Directed by Siddharth Anand, the
spy thriller, also featuring John Abraham and Deepika Padukone, has boosted Bollywood\xe2\x80\x99s flailing confidence
after a rough 2022. Pathaan is going strong in its second week, and is expected to have earned around Rs 6.7 crore on day
15 of release, according to the industry tracker Sacnilk. This brings the total domestic collection of the film to Rs
452.9 crore nett approximately. The film even managed to pass the second Monday test with flying colours, earning Rs 15.7
crore, and earned Rs 7.75 crore on Tuesday. Pathaan has already broken KGF: Chapter 2\xe2\x80\x99s record of Rs 434 crore
in the Hindi market.',
 b'For Bharat, who was a ball boy in 2005, when MS Dhoni announced himself on the international stage by hammering a
hapless Pakistan for 148 at Visakhapatnam, it was a dream come true.\n',
 b'2023 is the year where \xe2\x80\x9cArtificial Intelligence\xe2\x80\x9d or AI has dominated the discourse. Much of this
has been dominated by OpenAI\xe2\x80\x99s ChatGPT, the chatbot which has gone viral since it launched in November last
year, and already has over 100 million users. Microsoft has also gone all in on AI and invested $10 billion in OpenAI
along with announcing a new version of Bing which will be integrated with the start-up\xe2\x80\x99s AI technology. In
response to this Google has also announced its own chatbot called Bard, which has had a rough start thanks to one
incorrect answer. Let\xe2\x80\x99s take a look at key developments in AI news in the past few weeks.',
 b'Between 2014 and 2018, over 60% of IIT-Bombay graduates took up jobs in sectors not related to their branches of study,
a trend seen across all disciplines except Computer Science and\xc2\xa0 Engineering (CSE) and Electrical Engineering (EE),
according to a study by a group of researchers from the Centre for Policy Studies, IIT-Bombay.',
 b'The reason the RBI has stuck to the hawkish stance could lie in its outlook for India\xe2\x80\x99s economic growth and
inflation in 2023-24. The central bank expects the GDP to grow by 6.4%, but the growth rate to slow in every successive
quarter through the year; and for retail inflation to not fall below 5% in any quarter.',
 b'Under the protocol, Northern Ireland remains in the EU single market, and trade-and-customs inspections of goods coming
from Great Britain take place at its ports along the Irish Sea.',
 b"According to Congress sources, Singh's office asked the party to change his seat as it was difficult for him to walk to
the front row. The party then arranged for him to be seated in the back row, near the aisle.",
 b"Replying to the Motion of Thanks to the President's Address in Rajya Sabha, Prime Minister Narendra\nModi Thursday hit
out at the Congress government and said former Prime Minister Indira Gandhi used\nArticle 356 of the Constitution 50 times
to dismiss elected state governments.\n\n\x0c",
 b'Turkey Earthquake News Live Updates, February 09: Rescue efforts continued on Thursday as the\ndeath toll crossed
16,000, but hopes of finding survivors diminished on the fourth day of rescue workers.\nAs per experts, though people can
survive in the rubble for a week, the first 72 hours of the earthquake are\ncritical.\n\n\x0c',
 b'People above 14 years of age need to consume 2.4\nmicrograms (mcg) every day. However, the requirement\nchanges with
the calorie our body needs at a particular time.\nFor instance, a pregnant woman needs to ensure an intake of\n2.6 mcg,
and a lactating woman 2.8 mcg daily, says Dr\nSuranjit Chatterjee, Senior Consultant , Internal Medicine,\nIndraprastha
Apollo Hospital, New Delhi\n\n\x0c']
```

```
for i in range(len(l)):
    l[i]=str(l[i])
preprocessed_text=preprocess(l)
preprocessed_text
```



```
['bshah rukh khanxexx pathaan continu demolish record box offic direct siddharth anand spi thriller also featur john
abraham deepika padukon ha boost bollywoodxexx flail confid rough pathaan go strong second week expect earn around r crore
day releas accord industri tracker sacnilk thi bring total domest collect film r crore nett approxim film even manag pa
second monday test fli colour earn r crore earn r crore tuesday pathaan ha alreadi broken kgf chapter xexx record r crore
hindi market',
 'bfor bharat wa ball boy dhoni announc intern stage hammer hapless pakistan visakhapatnam wa dream come truen',
 'b year xexxcartifici intelligencexexxd ai ha domin discours much thi ha domin openaixexx chatgpt chatbot ha gone viral
sinc launch novemb last year alreadi ha million user microsoft ha also gone ai invest billion openai along announc new
version bing integr startupxexx ai technolog respons thi googl ha also announc chatbot call bard ha rough start thank one
incorrect answer letxexx take look key develop ai news past week',
 'bbetween iitbombay graduat took job sector relat branch studi trend seen across disciplin except comput scienc andxcxa
engin cse electr engin ee accord studi group research centr polici studi iitbombay',
 'bthe reason rbi ha stuck hawkish stanc could lie outlook indiaxexx econom growth inflat central bank expect gdp grow
growth rate slow everi success quarter year retail inflat fall ani quarter',
 'bunder protocol northern ireland remain eu singl market tradeandcustom inspect good come great britain take place port
along irish sea',
 'baccord congress sourc singh offic ask parti chang hi seat wa difficult walk front row parti arrang seat back row near
aisl',
 'brepli motion thank presid address rajya sabha prime minist narendranmodi thursday hit congress govern said former prime
minist indira gandhi usednarticl constitut time dismiss elect state governmentsnnxc',
 'bturkey earthquak news live updat februari rescu effort continu thursday thendeath toll cross hope find survivor
diminish fourth day rescu workersna per expert though peopl surviv rubbl week first hour earthquak arencriticalnnxc',
 'bpeopl abov year age need consum nmicrogram mcg everi day howev requirementnchang calori bodi need particular timenfor
instanc pregnant woman need ensur intak ofn mcg lactat woman mcg daili say drnsuranjit chatterje senior consult intern
medicinenindraprastha apollo hospit new delhinnxc']
```



```
generateInvertedIndexDict(preprocessed_text)
```

Output exceeds the size limit. Open the full output data in a text editor

```
{'bshah': {0},
 'rukh': {0},
 'khanxexx': {0},
 'pathaan': {0},
 'continu': {0, 8},
 'demolish': {0},
 'record': {0},
 'box': {0},
 'offic': {0, 6},
 'direct': {0},
 'siddharth': {0},
 'anand': {0},
 'spi': {0},
 'thriller': {0},
 'also': {0, 2},
 'featur': {0},
 'john': {0},
 'abraham': {0},
 'deepika': {0},
 'padukon': {0},
 'ha': {0, 2, 4},
 'boost': {0},
 'bollywoodxexx': {0},
 'flail': {0},
 'confid': {0},
 'rough': {0, 2},
 'go': {0},
 'strong': {0},
 'second': {0},
 'week': {0, 2, 8},
 'expect': {0, 4},
 'earn': {0},
 'around': {0},
 'r': {0},
 'crore': {0},
 'day': {0, 8, 9},
 'releas': {0},
 'accord': {0, 3},
 'industri': {0},
 'tracker': {0},
 'sacnilk': {0},
 'thi': {0, 2},
 'bring': {0},
 'total': {0},
```

```
 'domest': {0},
...
 'consult': {9},
 'medicinenindraprastha': {9},
 'apollo': {9},
 'hospit': {9},
 'delhinnxc': {9}}
```